Università di Cassino
Corso di Statistica 1
Esercitazione del 29/10/2007
Dott. Alfonso Piscitelli

Esercizio 1

Il seguente *data set* riporta la rilevazione di alcuni caratteri su un collettivo di 20 soggetti.

Soggetto	Sesso	Età	Reddito (Migliaia di €)	Titolo di studio	Nucleo familiare	Statura (cm)	Colore degli occhi
1	M	22	0,7	Diploma	3	173	NERO
2	F	18	0,2	Lic. Media	4	168	MARRONE
3	F	34	1,6	Diploma	2	165	MARRONE
4	M	42	2,5	Laurea	5	180	NERO
5	F	50	3,2	Diploma	3	163	AZZURRO
6	F	12	0,1	Lic. Elementare	4	160	NERO
7	M	46	3,8	Lic. Media	4	177	MARRONE
8	M	72	1,3	Nessun Titolo	2	164	VERDE
9	F	27	1,2	Laurea	3	158	AZZURRO
10	F	48	1,7	Lic. Media	5	170	NERO
11	F	35	1,9	Laurea	1	167	NERO
12	M	84	0,8	Nessun Titolo	1	159	MARRONE
13	F	21	0,4	Diploma	5	174	AZZURRO
14	F	44	1,8	Diploma	4	164	VERDE
15	M	56	1,9	Lic. Media	2	177	NERO
16	F	58	3,2	Lic. Media	3	172	NERO
17	F	37	2,1	Diploma	1	166	MARRONE
18	F	16	0,1	Lic. Media	4	160	MARRONE
19	M	73	1,6	Lic. Elementare	2	170	AZZURRO
20	M	64	2,2	Lic. Elementare	3	184	VERDE

a) Costruire la distribuzione di frequenza per il carattere **Età** suddividendo la distribuzione in 4 classi equiampie e determinare il valore del secondo decile, del settantesimo e novantesimo percentile.
b) Costruire la distribuzione doppia di frequenza per i caratteri **Sesso** e **Età**, utilizzando per questo ultimo carattere la suddivisione in classi operata precedentemente. Verificare la proprietà associativa della media.
c) Determinare la differenza interquartile per il carattere **Nucleo familiare** a partire sia dalla successione di valori sia dalla distribuzione di frequenze.
d) Si calcoli l'indice di Eterogeneità di Gini per il carattere **Colore degli occhi**.
e) Determinare l'indice di dispersione D per il carattere **Titolo di studio**.

Soluzioni

a) La distribuzione in classi di frequenza del carattere **Età**, è:

Età	d_i	n_i	N_i	f_i	F_i
12|-30	18	6	6	0,3	0,3
30|-48	18	6	12	0,3	0,6
48|-66	18	5	17	0,25	0,85
66|-|84	18	3	20	0,15	1,0
Tot:		**20**		**1**	

Prima di passare al calcolo del terzo decile, per dati in classi, bisogna evidenziare la classe in cui è presente il secondo decile. La classe del secondo decile è quella associata alla prima frequenza cumulata relativa che supera il valore di 0,20.

. ⇒ Classe D_2=12|-30 [in cui il valore della x associato alla prima frequenza relativa cumulata è maggiore di 0,20].

Quindi, il secondo decile sarà:

$$D_2 = l_d + \frac{\frac{\sum n_i}{10}(2) - N_{d-1}}{n_d} d_d$$

dove:

l_d =limite inferiore della classe del secondo decile;

N_{d-1} =frequenza cumulata associata alla classe precedente a quella del secondo decile;

n_d =frequenza assoluta della classe del secondo decile;

d_d =ampiezza della classe del secondo decile;

$$D_2 = 12 + \frac{4-0}{6}18 = 12 + 12 = 24$$

Le classi del settantesimo e del novantesimo percentile, si individuano in corrispondenza delle rispettive frequenze cumulate relative. Avremo quindi che:

. ⇒ Classe C_{70}=48|-66 [valore della x associato alla prima frequenza relativa cumulata maggiore di 0,70].

. ⇒ Classe C_{90}=66|-|84 [valore della x associato alla prima frequenza relativa cumulata maggiore di 0,90].

Quindi, il settantesimo percentile sarà:

$$C_{70} = l_{C70} + \frac{\frac{\sum n_i}{100}(70) - N_{C70-1}}{n_{C70}} d_{C70}$$

dove:

$l_{C_{70}}$ =limite inferiore della classe C_{70};

$N_{C_{70}-1}$ =frequenza cumulata associata alla classe precedente a quella di C_{70};

$n_{C_{70}}$ =frequenza assoluta della classe C_{70};

$d_{C_{70}}$ =ampiezza della classe C_{70};

$$C_{70} = 48 + \frac{14-12}{5}18 = 48 + 7,2 = 55,2$$

Quindi, il novantesimo percentile sarà:

$$C_{90} = l_{C90} + \frac{\frac{\sum n_i}{100}(90) - N_{C90-1}}{n_{C90}} d_{C90}$$

dove:

$l_{C_{90}}$ =limite inferiore della classe C_{90};

$N_{C_{90}-1}$ =frequenza cumulata associata alla classe precedente a quella di C_{90};

$n_{C_{90}}$ =frequenza assoluta della classe C_{90};

$d_{C_{90}}$ =ampiezza della classe C_{90};

$$C_{90} = 66 + \frac{18-17}{3}18 = 66 + 6 = 72$$

b) Per rappresentare la distribuzione doppia di frequenze dei due caratteri **Sesso** e **Età** (suddiviso in classi) è necessario costruirsi una tabella a doppia entrata che ha per righe le due modalità della variabile **Sesso** {Maschio, Femmina} e per colonne le 4 classi in cui è stata suddivisa la variabile **Età** {12|-30; 30|-48; 48|-66; 66|-|84 }.

Si precisa che la stessa informazione si avrebbe da una tabella che ha per righe le classi della variabile **Età** e per colonne le modalità della variabile **Sesso**.
La distribuzione doppia di frequenza delle due variabili è:

	12|-30	30|-48	48|-66	66|-|84	Tot:
Maschio	1	2	2	3	8
Femmina	5	4	3	0	12
Tot:	6	6	5	3	20

In questa tabella, il calcolo della media aritmetica della variabile **Età** sarà:

$$\mu = \frac{1}{N}\sum_{i=1}^{C} x_i^c * n_i$$

Ricordando che i valori centrali delle classi sono rispettivamente: 21; 39; 57; 75.

$$\mu = \frac{(21*6)+(39*6)+(57*5)+(75*3)}{20} = \frac{870}{20} = 43{,}5$$

Per poter verificare la proprietà associativa della media bisogna ricorrere alle distribuzioni di frequenza condizionate. Le due distribuzioni di frequenza condizionate della variabile **Età** sono:

	12|-30	30|-48	48|-66	66|-|84	Tot:
Maschio	1	2	2	3	**8**

	12|-30	30|-48	48|-66	66|-|84	Tot:
Femmina	5	4	3	0	**12**

Per ognuna delle tabelle si calcola la media della distribuzione condizionata delle variabile **Età** rispetto alle variabile **Sesso**:

$$\mu_{Età|M} = \frac{(21*1)+(39*2)+(57*2)+(75*3)}{8} = \frac{438}{8} = 54{,}75$$

L'Età media dei Maschi è 54,75

$$\mu_{Età|F} = \frac{(21*5)+(39*4)+(57*3)+(75*0)}{12} = \frac{432}{12} = 36$$

L'Età media delle Femmine Maschi è 36

La proprietà associativa della Media afferma che la media delle medie condizionate ponderata per la numerosità del gruppo è uguale alla media generale. Essa è verificata dalla seguente uguaglianza:

$$\frac{\sum_{i=1}^{G} \mu_i * n_i}{\sum n_i} = \mu$$

dove G= numero di gruppi.

Nel nostro caso G=2 e la media delle medie condizionate è:

$$\mu = \frac{\sum_{i=1}^{2} \mu_i * n_i}{\sum n_i} = \frac{(54{,}57 * 8) + (36 * 12)}{20} = \frac{870}{20} = 43{,}5$$

c) La successione dei valori ordinati in senso non decrescente e la corrispondente distribuzione di frequenza della variabile **Nucleo familiare** sono le seguenti:

Soggetto	Posizione	Nucleo familiare
11	1	1
12	2	1
17	3	1
3	4	2
8	5	2
15	6	2
19	7	2
1	8	3
5	9	3
9	10	3
16	11	3
20	12	3
2	13	4
6	14	4
7	15	4
14	16	4
18	17	4
4	18	5
10	19	5
13	20	5

Nucleo familiare	n_i	f_i	F_i
1	3	0,15	0,15
2	4	0,2	0,35
3	5	0,25	0,60
4	5	0,25	0,85
5	3	0,15	1
Tot:	**20**	**1**	

Il primo quartile corrisponde a quel valore del carattere X che lascia alla sua sinistra il 25% delle osservazioni e alla sua destra il 75%.
Il terzo quartile corrisponde a quel valore del carattere X che lascia alla sua sinistra il 75% delle osservazioni e alla sua destra il rimanente 25%.

$$Q_1 = \frac{X_{\frac{N}{4}} + X_{\frac{N}{4}+1}}{2} = \frac{X_5 + X_6}{2} = \frac{2+2}{2} = 2$$

$$Q_3 = \frac{X_{3*\frac{N}{4}} + X_{3*\frac{N}{4}+1}}{2} = \frac{X_{15} + X_{16}}{2} = \frac{4+4}{2} = 4$$

Nel caso delle distribuzioni di frequenza semplice, invece, i quartili vengono individuati facendo riferimento alle frequenze cumulate o alle frequenze relative cumulate. In questo caso:

. ⇒ il primo quartile è quel valore della x associato alla prima frequenza relativa cumulata maggiore di 0,25. [Q_1=2]

. ⇒ il terzo quartile è quel valore della x associato alla prima frequenza relativa cumulata maggiore di 0,75. [Q_3=4].

Si definisce differenza interquartile la differenza tra il terzo e il primo quartile.

IQR= Q_3 - Q_1 = 4-2=2

Questa quantità contiene il 50% "centrale" delle osservazioni.

d) Nel caso di variabili qualitative la variabilità del carattere è espressa in termini di **mutabilità**, definita come l'attitudine di un carattere ad assumere differenti modalità qualitative.

Quando tutte le unità statistiche assumono la stessa modalità, si ha una **perfetta omogeneità**. (minima eterogeneità)

Quando le modalità del carattere hanno tutte la stessa frequenza assoluta o relativa, si ha la **massima disomogeneità**.

L'Eterogeneità misura la variabilità delle frequenze delle *k* modalità del carattere.

L'**Indice di Eterogeneità** (**G**) di Gini si basa sulle frequenze relative.

$$G = 1 - \sum_{i=1}^{k} f_i^{\,2}$$

Si tratta di un indice relativo che varia tra $0 \le G \le 1 - \frac{1}{k}$

G=0 si ha la minima eterogeneità.

$G = 1 - \frac{1}{k}$ si ha la massima eterogeneità.

La distribuzione di frequenza della variabile **Colore degli occhi** è:

Colore degli occhi	**n_i**	**f_i**
Nero	7	0,35
Marrone	6	0,3
Azzurro	4	0,2
Verde	3	0,15
Tot:	**20**	**1**

Colore degli occhi	f_i	$(f_i)^2$
Nero	0,35	0,1225
Marrone	0,3	0,09
Azzurro	0,2	0,04
Verde	0,15	0,0225
Tot:	**1**	**0,275**

Quindi **G** sarà:

$$G = 1 - \sum_{i=1}^{k} f_i^{\,2} = 1 - 0{,}275 = 0{,}725$$

Volendo normalizzare G si divide il valore ottenuto per il suo massimo $1-\frac{1}{k}$ ottenendo così G*

$$G^* = \frac{G \cdot k}{k-1} = 0{,}9667$$

Si può dire che siamo molto vicini al caso di massima eterogeneità.

e) Per poter effettuare il calcolo dell'indice di dispersione **D** per il carattere **Titolo di studio**, bisogna partire dalla distribuzione di frequenze

Titolo di studio	n_i	f_i	F_i
Nessun Titolo	2	0,10	0,10
Lic. Elementare	3	0,15	0,25
Lic. Media	6	0,30	0,55
Diploma	6	0,30	0,85
Laurea	3	0,15	1
Tot:	**20**	**1**	

L'indice di dispersione **D**, a differenza di altri indici di omogeneità **/** eterogeneità utilizzati per le variabili qualitative nominali, consente di utilizzare l'ulteriore informazione detenuta dalle variabili qualitative ordinali, ovvero la possibilità di ordinarne le modalità.

$$D = 2\sum_{i=1}^{k-1} F_i(1 - F_i)$$

Titolo di studio	n_i	f_i	F_i	$(1- F_i)$	$F_i *(1- F_i)$
Nessun Titolo	2	0,10	0,10	0,90	0,0900
Lic. Elementare	3	0,15	0,25	0,75	0,1875
Lic. Media	6	0,30	0,55	0,45	0,2475
Diploma	6	0,30	0,85	0,15	0,1275
Laurea	3	0,15	1	0	
Tot:	**20**	**1**			**0,6525**

D=2*0,6525=1,305

Sapendo che il valore massimo che può assumere l'indice nel caso di numerosità pari è:

$$D_{MAX} = \frac{K-1}{2} = \frac{5-1}{2} = 2$$

è possibile calcolare l'indice D normalizzato tra [0 - 1].

$$D_{[0-1]} = \frac{D}{D_{MAX}} = \frac{1,305}{2} = 0,6525$$